Image-based screening of oral cancer via deep ensemble architecture

Marco Parola
Dept. of Information Engineering
University of Pisa
Largo L. Lazzarino 1, Pisa, Italy
marco.parola@ing.unipi.it
0000-0003-4871-4902

Gaetano La Mantia
Department Di.Chir.On.S
University of Palermo
Palermo, Italy
gaetano.lamantia@community.unipa.it
0000-0002-3135-7462

Federico Galatolo
Dept. of Information Engineering
University of Pisa
Largo L. Lazzarino 1, Pisa, Italy
federico.galatolo@unipi.it
0000-0001-7193-3754

Mario G.C.A. Cimino
Dept. of Information Engineering
University of Pisa
Largo L. Lazzarino 1, Pisa, Italy
mario.cimino@unipi.it
0000-0002-1031-1959

Giuseppina Campisi
Department Di.Chir.On.S
University of Palermo
Palermo, Italy
campisi@odonto.unipa.it
0000-0002-9443-0495

Olga Di Fede
Department Di.Chir.On.S
University of Palermo
Palermo, Italy
odifede@odonto.unipa.it
0000-0002-5562-7420

***Abstract*—Oral squamous cell carcinoma (OSCC) is a significant health issue in the oral cancer domain; a screening tool for timely and accurate diagnosis is essential for effective treatment planning and prognosis in patients' life expectancy. In this paper, we address the problem of object detection and classification in the context of OSCC, by presenting a comparative analysis of three state-of-the-art architecture: YOLO, FasterRCNN, and DETR. We propose a deep learning ensemble model to address both object detection and classification problem leveraging the strengths of individual models to achieve higher performance than single models. The proposed architecture was evaluated on a real-world dataset developed by experienced clinicians who manually labeled individual photographic images, producing a benchmark dataset. Results from our comparative analysis demonstrates the ensemble detection model achieves superior performance compared to the individual models, outperforming the average value of the individual models' map@50 metric by 24% and the value of the map@95-50 metric by 44%.**

***Index Terms*—Oral cancer, Oral squamous cell carcinoma, YOLO, DETR, Faster R-CNN, Ensemble, Lesion detection.**

I. Introduction

Oral squamous cell carcinoma (OSCC) is a severe public health concern in low-income and emerging nations [1].

Early detection of OSCC lesions is critical to the successful treatment of this deadly disease, which is classified in the category of oral cavity cancers. Under this premise, the detection of such lesions is a challenging task requiring experienced health professionals, and the accuracy of their assessments can be affected by various factors, such as the size, location, and morphology of the lesion. Therefore, there is growing interest in the development of computer-aided screening systems that can support oral health care providers in the detection of OSCC lesions with high sensitivity and specificity [2].

Over the past decade, Deep learning (DL) techniques have been progressively introduced to address various challenges associated with medical image analysis [3]. One of the main advantages of DL over traditional techniques lies in its ability to automatically learn representations directly from raw data [4][5]. Unlike traditional methods, which rely on hand-created features, DL models are able to autonomously extract relevant features and patterns from medical images, enabling analysis that is less dependent on the parameterization algorithm stages. DL has shown great potential in healthcare applications, including medical image analysis, diagnosis and treatment. DL-based tools are being progressively integrated in healthcare processes as they are able to analyze large amounts of data and identify subtle patterns sometimes difficult for humans to detect [6].

Numerous studies have focused on the design and development of deep learning systems capable of automating lesion detection and classification in the oral cavity. These systems rely on various diagnostic techniques and expensive image acquisition machines such as laser confocal endomicroscopy, autofluorescence imaging, hyperspectral imaging, optical coherence tomography. [7–9], which typically involve invasive techniques and demand substantial time and personnel expertise to acquire such data. However, the development of technology in recent years has opened up new possibilities for effective and noninvasive diagnostic methods. This research paper aims to highlight a crucial element in the specific context of oral cancer: the use of photographic images. The wide availability of cameras, both as stand-alone devices and integrated into smartphones, has greatly improved the ease of photographically documenting oral lesions in the medical setting [10], enabling large-scale screening relying on more widespread and frequent testing in the population to detect cancer at early stages.

In this paper, we present a comprehensive evaluation and

comparison of three state-of-the-art DL architectures for image detection and classification: You Only Look Once (YOLO) version 8 [11], Faster Region Convolutional Neural Network (FasterRCNN), implemented in the Detectron2 framework [12], and DEtection TRansformer (DETR) [13]. Our objective is to design a screening tool that addresses both object detection and classification tasks in a unified manner.

The paper is structured as follows. The material and methodology are described in Section 2, while the case study and experiment results are discussed in Section 3 and Section 4, respectively. Finally, Section 5 draws conclusions and outlines future research possibilities.

II. Related work

Several researches in the literature have contributed to enhance the classification and identification of OSCC in the context of oral cancer. These studies have used object detection models and deep CNN to analyze oral photographic data.

In 2020, Welikala et al. proposed work on automatic detection and classification of oral cavity injuries using DL [14]. Starting with 2155 images of the oral cavity of 1085 individuals, containing images with and without lesions, produced a dataset using the composite annotation method, whereby the annotation of an image is generated from multiple annotations of the same image made by different clinicians. Two deep learning-based computer vision approaches for automated detection and classification of oral lesions for early detection of oral cancer were then evaluated: image classification with ResNet-101 and object detection with FasterRCNN. Image classification achieved an F1 score of 0.87, while object detection achieved an F1 score of 0.41. The F1 metrics were defined differently for classification and detection.

In 2021 Warin et al. presented a study aiming to design a CNN-based screening tool for the classification and detection of oral cancer [15]. The authors prepared a dataset consisting of 700 clinical photographs, divided into 350 images of oral squamous cell carcinoma and 350 images of healthy oral mucosa. DenseNet121 and FasterRCNN were introduced for classification and detection tasks, respectively. DenseNet121 achieved 0.99 accuracy, 1.0 recall and 0.99 F1 score in the classification task. While the detection performance of a FasterRCNN model achieved 0.76 accuracy, 0.82 recall and 0.79 F1 score in the lesion detection task. Again, the evaluation metrics were defined differently for classification and detection. In observing the previous results, it is important to critically analyze the validity and reliability. In particular, there seems to be some potential errors or discrepancies in the classification evaluation. In fact, a 99 percent accuracy value is very high and in real-world scenarios it could indicate potential overfitting, data loss, or an unrealistic evaluation setup.

Another important study in which Transformer architecture was introduced in the context of oral cancer was conducted by Flugge et al. in [16] in 2023. The authors solved an automatic OSCC classification problem in clinical photographs without performing the detection task. They used a DL approach based on Swin-Transformer, trained on 1124 images and tested an additional 141. The proposed method achieved a classification accuracy of 0.986. The study authors pointed out that the lack of experience and training of primary operators may lead to diagnostic delays and, as a result, more extensive surgical procedures with prolonged hospitalization and lower survival rates. They also pointed out the limitations of their study related to the single-center design (data are collected from a single site or location); in order to develop accurate DL models in the healthcare environment, the data source must be very large and varied. This work was reported in the literature review, despite the authors not performing the detection task, because it is the first to have introduced the transformer in the context of oral cancer

III. Methodology

The identification of the oral cavity part affected by a lesion can be formulated as a supervised learning problem: the object detection. Let's consider a set of labeled training images denoted by $\mathscr{D}$, where each image I is associated with a set of bounding boxes B_I and their corresponding class labels L_I both provided by a domain expert; each bounding box B_i is represented by four parameters: (x, y, w, h), where (x, y) denotes the coordinates of the top-left corner, and (w, h) denote the width and height of the bounding box, respectively. The class label c_i represents the class of the object enclosed by the bounding box.

The goal is to learn a mapping function $f_\theta(I)$, typically implemented using a DL model, that accurately predict the bounding boxes and class labels of objects in test images, where θ represents the learnable parameters of the model.

The learning algorithm optimizes the model parameters θ by minimizing the loss function $\mathscr{L}$ over the labeled training images as shown in the Equation 1. $\mathscr{L}$ is the result of two separate contributions: the localization loss $\mathscr{L}_{\text{loc}}$ measuring the bounding box positions, and the classification loss $\mathscr{L}_{\text{cls}}$ evaluating the classification accuracy, shown in equations 2 and 3, respectively. The specific formulations for L_{loc} and L_{cls} depend on the chosen methods and architectures as well as the desired properties of the model. Common choices include smooth L1 loss for localization and cross-entropy loss for classification.

$$\mathscr{L}(f_\theta(I), B_I, L_I) = \lambda_{\text{loc}}\mathscr{L}_{\text{loc}} + \lambda_{\text{cls}}\mathscr{L}_{\text{cls}} \tag{1}$$

$$\mathscr{L}_{\text{loc}} = \sum_{i \in \mathscr{B}_{\text{pos}}} L_{\text{loc}}(B_I, B_I*) \tag{2}$$

$$\mathscr{L}_{\text{cls}} = \sum_{i \in \mathscr{B}_{\text{pos}}} L_{\text{cls}}(c_I, c_I*) \tag{3}$$

where λ_{loc} and λ_{cls} are weights to balance the contributions of the different loss components. $\mathscr{B}_{\text{pos}}$ represents the set of indices of positive samples. Finally, B_i and c_i represent the predicted bounding box and class label for the i-th object, while B_i* and c_i* represent the ground truth bounding box and class label for the i-th object.

The comparison of different object detection models was proposed by introducing the metrics below [17]. Intersection over Union (IoU) is a metric based on the Jaccard coefficient, measuring the overlap between the predicted bounding box and the ground truth bounding box. It is calculated by dividing the area of intersection by the area of union between the two bounding boxes.

Mean Average Precision at a single threshold (mAP@th) is a popular metric to evaluate the performance of object detection models. It computes the well-known average precision (AP) metric at a specific threshold value of the IoU metric for each class and then takes the mean across all classes.

An analogous metric is mean Average Precision between thresholds th1 and th2 (mAP@th_1-th_2). This metric is similar to mAP@th, but instead of considering a single threshold, it computes the average precision across a range of thresholds, specified by the upper and lower thresholds, th_1 and th_2 respectively. Since, mAP@th_1-th_2 considers a wider range of confidence thresholds, it provides a more balanced assessment and less affected by single-threshold model performance.

Equations 4 and 5 present such metrics with the actual parameter values used during the experiment phase: mAP@50 and mAP@95-50, respectively.

$$mAP@50 = \frac{1}{N}\sum_{i=1}^{N} AP@50_i \tag{4}$$

$$mAP@95-50 = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=0}^{9} AP@(50+5j)_i \tag{5}$$

where N is the number of classes.

A. Deep learning models

In this paper, we have introduced three state-of-the-art architectures that will be described in this subsection: FasterRCNN, YOLO, and DETR.

First of all, FasterRCNN is a DL architecture for object detection in images. The model is based on two main components: a Region Proposal Network (RPN) and a region-based detector. The RPN is responsible for generating a set of ROIs potentially containing objects. These regions are then fed into the region-based detector, which exploits the convolutional backbone to extract features from the image. The extracted features are then used to classify objects and refine their bounding boxes. The FasterRCNN differs from previous designs in that it generates bounding boxes using RPN rather than algorithm-based selective search. This enables a more efficient and accurate region proposal, which in turn improves overall detection performance.

The You Only Look Once (YOLO) architecture is the second DL-based model for object detection that we introduced. Unlike other architectures, which divide the image into small regions and analyze them separately, YOLO examines the entire image in a single step, detecting objects and providing a class of membership among a predetermined set for each of them.

A key component of YOLO is nonmaximal suppression (NMS); after YOLO has predicted multiple bounding boxes around the detected objects, NMS filters out the redundant bounding boxes to improve the accuracy of object location. The procedure begins by sorting the predicted bounding boxes by their confidence score. The box with the highest score is selected as the reference. Then, each subsequent box is compared with the reference box using the IoU metric.If the IoU value exceeds a predefined threshold, the overlapping box is considered a duplicate and deleted. However, if the IoU value is below the threshold, the box is kept as a separate detection. By applying NMS, YOLO effectively eliminates duplicate detections, resulting in a more accurate representation of object positions.

Finally, DEtection TRansformer is an object detection model based on a transformer architecture, originally developed for natural language processing and later applied to computer vision [18].

DETR consists of a convolutional backbone followed by an encoder-decoder transformer that can be trained end-to-end for object detection. The encoder is applied to the spatial features extracted from the input image by the backbone; then, the decoder maps these features to generate output bounding boxes. A key element of DETR is the use of multi-headed self-attention mechanisms within the transformer architecture. These self-attention heads allow DETR to capture global context information and model the relationships between different object instances. Moreover, the multi-headed self-attention mechanism allows DETR to handle different object scales and aspect ratios effectively. The heads learn to attend to different spatial regions, capturing both fine-grained details and higher-level context.

B. Ensemble architecture

In order to address the lesion detection problem in the context of oral cancer, we present a proposed DL architecture based on an ensemble of three state-of-the-art object detection models: YOLO, FasterRCNN, and DETR.

Fig. 1 presents an overview of the overall ensemble architecture. It consists of three individual object detection models, represented as orange rectangles; each taking the same input image and producing an output of the same shape, which includes bounding box coordinates and corresponding labels. The outputs from these models are then fed into an aggregator module, that combines them to produce the final predictions for both the label and bounding box coordinates. It consists of two submodules: (i) weighed average voting and (ii) window fusion, in Figure 1 depicted in green and red, respectively.

In the average voting submodule, the outputs of the three models are multiplied by a weight and summed to construct a vector equal in size to the number of classes, in which each model contributes to increasing the value of the corresponding class by an amount equal to its weight for that class. Finally, this vector is transformed into a probability vector using the softmax function and the final label $label_E$ is obtain by the

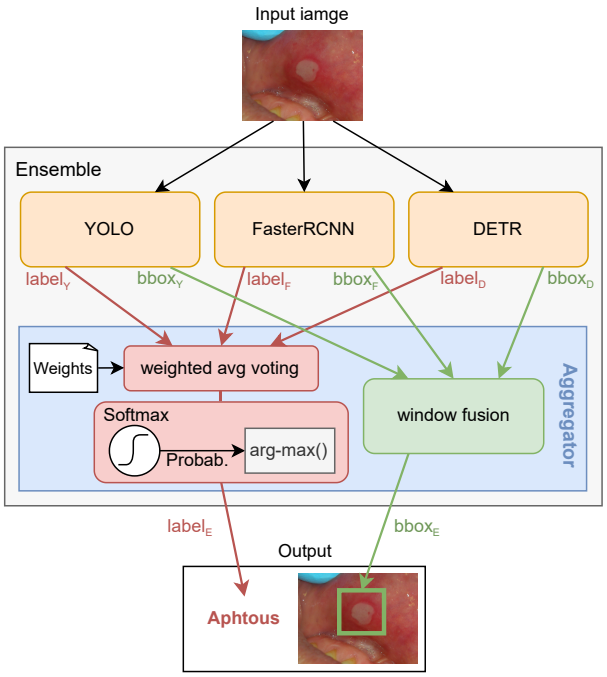

Fig. 1. Ensemble architecture.

$arg-max$ function; Equation 6 illustrates the above from a mathematical point of view.

$$label_E = \textit{arg-max}\left(\textit{Softmax}\sum_{i\in M} w_i X_i\right) \tag{6}$$

where $M = \{Y, F, D\}$ is the set of three models[1]; w_i represents the vector weights assigned to the i-th model; X_i denotes its the categorical output label.

The window fusion submodule is responsible for predicting the final bounding box coordinates. Specifically, given a set $\mathscr{B} = \{B_{Y1}, ..., B_{Yn}, B_{F1}, ..., B_{Fm}, B_{D1}, ..., B_{Dl}\}$ of bounding boxes provided by the three different models for an input image, the set of the ensemble model predicted bounding boxes $\mathscr{B}_E = \{B_{E1}, \ldots, B_{Ep}\}$ is generated by evaluating the overlapping between different boxes. Specifically, each bounding box $B_{E\alpha}$ is generated from a subset $\hat{\mathscr{B}} \subseteq \mathscr{B}$ by performing the intersection between each element, as shown in Equation 6; where $\hat{\mathscr{B}}$ is composed of all the elements such that $\forall B_\beta, B_\gamma \in \mathscr{B}, IoU(B_\beta, B_\gamma) > th, \gamma \neq \beta$.

Additionally, in order to build a robust ensemble model, we pose the following constraint $\exists\ B_{ij}, B_{hk} \mid i \neq h; i, h \in M$, meaning that the ensemble bounding box is computed between at least two bounding boxes predicted by different models. Hence, if a lesion is identified only by one model, the ensemble considers it a false.

$$B_{E\alpha} = \cap_i B_i \qquad \forall B_i \in \hat{\mathscr{B}};\ \alpha = 1, ..., p \tag{7}$$

[1] $Y = YOLO, F = FasterRCNN, D = DETR, E = Ensemble.$

IV. Oral case study

Between 2021 and 2023, images of the oral cavity were collected from patients visiting the Oral Medicine Unit of the P. Giaccone University Hospital in Palermo, Italy. Images were captured with a smartphone camera or standard camera by oral medicine practitioners (i.e. dental hygienists, consultants and trainees), avoiding expensive imaging machines. Tab. I summarizes the number of acquired images by class.

TABLE I
Annotation's numerosity per class

Class	Samples
Aphthous	142
Neoplastic	144
Traumatic	142
All	428

We used the COCO Annotator annotation tool to annotate the images. A trained dentist manually annotated the lesions in the images. Each lesion was annotated with a bounding segment and a corresponding label, the bounding box being generated by the tool from the segment. The annotations were reviewed by a senior dentist to ensure consistency and accuracy.

V. Experiments and Results

The software for experiments has been implemented in Python, and publicly released on GitHub to guarantee repeatability and transparency [19]. The hardware resources used for the experiment include the Intel i7-1280P CPU, a Nvidia GeForce GTX 1650 GPU and 32 GiB of RAM. In order to enhance the robustness and generalization capabilities of our object detection model, we conducted a data augmentation phase prior to the experiment. This phase involved generating four augmented images for each original image in the dataset, along with the corresponding bounding boxes as labels. The augmentation techniques applied included geometric random transformations such as rotations (between -10 and 10 degree), translations both on x and y axis (between -10% and 10%), and scaling (90-110% of original), as well as photometric transformations such as adjustment in brightness, contrast (95-105% of original).

The tables II show the values of the performance evaluation metrics mAP@50 and mAP@95-50 obtained from YOLOv8, FasterRCNN, and DETR on the individual classes and on the whole dataset.

In Fig. V, we present some interesting cases to better understand how our ensemble model performs in different scenarios compared with the ground truth in the first column and the three individual models in the following three columns. The first case (a) represents an optimal scenario in which all individual models correctly detected and classified the oral lesion. Consequently, our ensemble model also performed well, accurately identifying and classifying the lesion. In the second case (b), we observed that although all the models accurately detected the lesion, only the DETR classified it

TABLE II
COMPARISON OF MAP METRICS BY CLASSES AMONG THE YOLOV8, FASTER R-CNN AND DETR MODELS.

Model	Class	MAP@50	MAP@50-95
YOLOv8	Neoplastic	.457	.196
	Aphthous	.314	.144
	Traumatic	.490	.208
	All	.421	.183
FasterRCNN	Neoplastic	.722	.288
	Aphthous	.279	.108
	Traumatic	.392	.153
	All	.464	.183
DETR	Neoplastic	.631	.296
	Aphthous	.303	.112
	Traumatic	.459	.172
	All	.465	.193

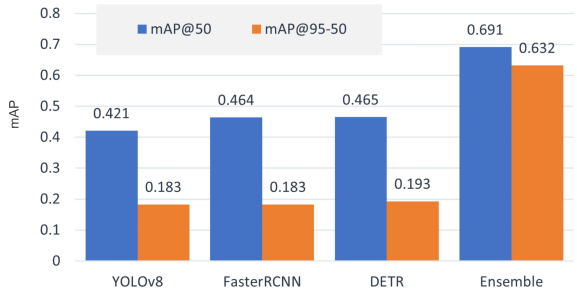

Fig. 2. Comparison of map metrics between YOLOv8, FasterRCNN, DETR and ensemble models on the test set.

correctly. Despite the fact just one out of three models provided a correct classification, our ensemble model was able to classify it correctly. This was achieved by assigning a higher weight to the "traumatic" label in the ensemble model compared to the sum of the "aphthous" YOLO and FasterRCNN weight labels. In the third case (c), which is also a common occurrence, two out of three models successfully detected and classified the neoplastic lesion. Consequently, our ensemble model computed the intersection between the two bounding boxes to determine the output. Also in this case, as in the previous one, we can observe how the weighted average voting ensures a consistent final prediction. The fourth case (d) proved problematic, as all models misclassified the prediction, leading to an incorrect ensemble prediction. The way to overcome this limitation is to enhance individual model performance with a larger dataset. In case (e), we encountered a situation where the DETR model predicted two bounding boxes, but only one of them was correct. According to our ensemble model logic, this prediction was discarded since it was unique to DETR and not present in YOLO and FasterRCNN and it was considered a false positive. Finally, in the last case (f), we observed a scenario where the ensemble model performed worse than a single model (DETR), which would have correctly detected the lesion. However, following the same reasoning as in the previous case, where only one model detected it, the ensemble model treated it as a false positive.

VI. CONCLUSIONS

In this study, we proposed an ensemble architecture for oral squamous cell carcinoma detection based on a set of three object detection models: YOLO, FasterRCNN and DETR. The goal was to exploit the strengths of these models and improve the overall performance of oral cancer detection.

The performance improvement achieved with the ensemble architecture is significant and confirms the effectiveness of combining multiple models. By aggregating the predictions of the three models, we were able to improve the robustness of the sensing system, resulting in improved overall performance.

The results of our experimental evaluation indicate that the ensemble architecture outperforms the individual models, especially when considering the mAP@50 metric. The ensemble architecture obtained a mAP@50 score of 0.69, significantly exceeding the average value of 0.45 obtained by the individual models. However, what really distinguishes the ensemble architecture is its performance on the mAP@95-50 metric. In this regard, it achieved a mAP@95-50 score of 0.63, while the average value of the individual models was only 0.19. These results highlight the robustness of the ensemble model to different levels of confidence in the detected objects.

It is important to note that adopting the ensemble architecture requires training three distinct models. This process requires more time and computational resources compared to using a single model. Although this requires initial effort, it can be considered an acceptable trade-off when considering the performance improvement achieved.

In addition, although the proposed architecture has demonstrated superior performance, periodic retraining of the models may be necessary to maintain optimal results. As new data become available or the distribution of oral cancer images evolves, retraining of the models may help to ensure their continued effectiveness. The frequency of retraining will depend on the specific requirements of the application and the availability of new data.

ACKNOWLEDGMENT

Work partially supported by: (i) the University of Pisa, in the framework of the PRA 2022 101 project "Decision Support Systems for territorial networks for managing ecosystem services"; (ii) the European Commission under the NextGenerationEU program, Partenariato Esteso PNRR PE1 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI"; (iii) the Italian Ministry of Education and Research (MIUR) in the framework of the FoReLab project (Departments of Excellence) and of the "Reasoning" project, PRIN 2020 LS Programme, Project number 2493 04-11-2021.

REFERENCES

[1] J. Musulin, D. Stifanic, A. Zulijani, S. B. Segota, I. Lorencin, N. Andjelic, and Z. Car, "Automated grading of oral squamous cell carcinoma into multiple classes using deep learning methods," in *2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 1–6, IEEE, 2021.

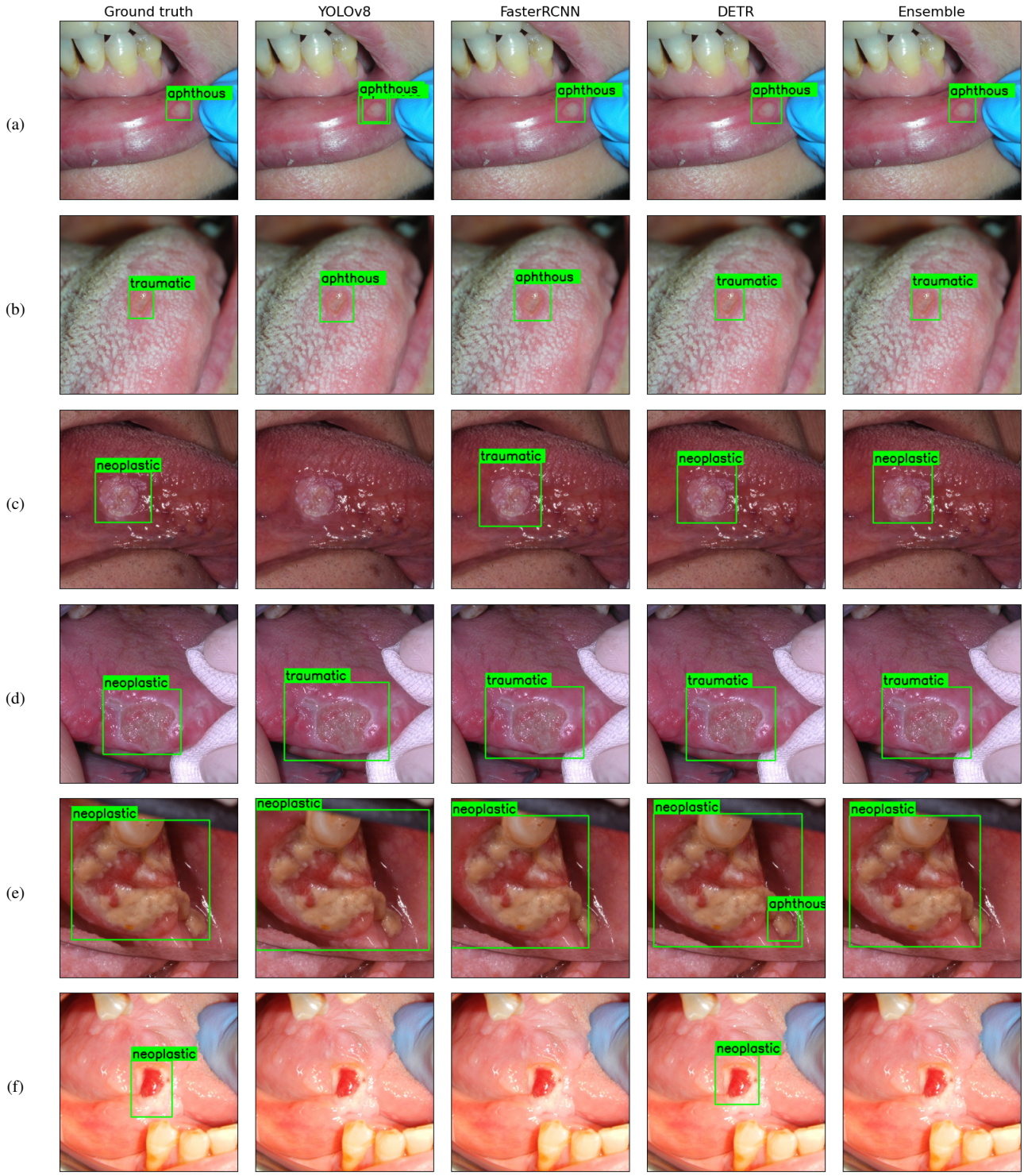

Fig. 3. Six examples of inference of the 4 models against the ground truth

[2] H.-P. Chan, L. M. Hadjiiski, and R. K. Samala, "Computer-aided diagnosis in the era of deep learning," *Medical physics*, vol. 47, no. 5, pp. e218–e227, 2020.

[3] D. Raimondo, A. Raffone, A. C. Aru, M. Giorgi, I. Giaquinto, E. Spagnolo, A. Travaglino, F. A. Galatolo, M. G. C. A. Cimino, J. Lenzi, G. Centini, L. Lazzeri, A. Mollo, R. Seracchioli, and P. Casadio, "Application of deep learning model in the sonographic diagnosis of uterine adenomyosis," *International Journal of Environmental Research and Public Health*, vol. 20, no. 3, 2023.

[4] F. Galatolo., M. Cimino., and E. Cogotti., "Tetim-eval: A novel curated evaluation data set for comparing text-to-image models," in *Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods - ICPRAM*, pp. 590–596, INSTICC, SciTePress, 2023.

[5] F. A. Galatolo., M. G. C. A. Cimino., and G. Vaglini., "Generating images from caption and vice versa via clip-guided generative latent space search," in *Proceedings of the International Conference on Image Processing and Vision Engineering - IMPROVE*, pp. 166–174, INSTICC, SciTePress, 2021.

[6] W. Walter, C. Pohlkamp, M. Meggendorfer, N. Nadarajah, W. Kern, C. Haferlach, and T. Haferlach, "Artificial intelligence in hematological diagnostics: Game changer or gadget?," *Blood Reviews*, p. 101019, 2022.

[7] H. M. Afify, K. K. Mohammed, and A. E. Hassanien, "Novel prediction model on oscc histopathological images via deep transfer learning combined with grad-cam interpretation," *Biomedical Signal Processing and Control*, vol. 83, p. 104704, 2023.

[8] M. Aubreville, C. Knipfer, N. Oetter, C. Jaremenko, E. Rodner, J. Denzler, C. Bohr, H. Neumann, F. Stelzle, and A. Maier, "Automatic classification of cancerous tissue in laserendomicroscopy images of the oral cavity using deep learning," *Scientific reports*, vol. 7, no. 1, p. 11979, 2017.

[9] E. Duran-Sierra, S. Cheng, R. Cuenca, B. Ahmed, J. Ji, V. V. Yakovlev, M. Martinez, M. Al-Khalil, H. Al-Enazi, Y.-S. L. Cheng, *et al.*, "Machine-learning assisted discrimination of precancerous and cancerous from healthy oral tissue based on multispectral autofluorescence lifetime imaging endoscopy," *Cancers*, vol. 13, no. 19, p. 4751, 2021.

[10] B. Hunt, A. J. Ruiz, and B. W. Pogue, "Smartphone-based imaging systems for medical applications: a critical review," *Journal of Biomedical Optics*, vol. 26, no. 4, pp. 040902–040902, 2021.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

[12] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2." https://github.com/facebookresearch/detectron2, 2019.

[13] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pp. 213–229, Springer, 2020.

[14] R. A. Welikala, P. Remagnino, J. H. Lim, C. S. Chan, S. Rajendran, T. G. Kallarakkal, R. B. Zain, R. D. Jayasinghe, J. Rimal, A. R. Kerr, *et al.*, "Automated detection and classification of oral lesions using deep learning for early detection of oral cancer," *IEEE Access*, vol. 8, pp. 132677–132693, 2020.

[15] K. Warin, W. Limprasert, S. Suebnukarn, S. Jinaporntham, and P. Jantana, "Automatic classification and detection of oral cancer in photographic images using deep learning algorithms," *Journal of Oral Pathology & Medicine*, vol. 50, no. 9, pp. 911–918, 2021.

[16] T. Flügge, R. Gaudin, A. Sabatakakis, D. Tröltzsch, M. Heiland, N. van Nistelrooij, and S. Vinayahalingam, "Detection of oral squamous cell carcinoma in clinical photographs using a vision transformer," *Scientific Reports*, vol. 13, no. 1, p. 2296, 2023.

[17] R. Padilla, S. L. Netto, and E. A. Da Silva, "A survey on performance metrics for object-detection algorithms," in *2020 international conference on systems, signals and image processing (IWSSIP)*, pp. 237–242, IEEE, 2020.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[19] M. Parola and M. Lorenzo, "Github oral detection code repository, https://github.com/marcoparola/detection_framework," 2023.